

Non-asymptotic Oracle Inequalities for the High-Dimensional Cox Regression via Lasso

Shengchun Kong and Bin Nan *

*Department of Biostatistics
University of Michigan
1420 Washington Heights
Ann Arbor, MI 48109-2029
e-mail: kongsc@umich.edu*

*bnan@umich.edu

Abstract: We consider the finite sample properties of the regularized high-dimensional Cox regression via lasso. Existing literature focuses on linear models or generalized linear models with Lipschitz loss functions, where the empirical risk functions are the summations of independent and identically distributed (iid) losses. The summands in the negative log partial likelihood function for censored survival data, however, are neither iid nor Lipschitz. We first approximate the negative log partial likelihood function by a sum of iid non-Lipschitz terms, then derive the non-asymptotic oracle inequalities for the lasso penalized Cox regression using pointwise arguments to tackle the difficulty caused by the lack of iid and Lipschitz property.

AMS 2000 subject classifications: 62N02.

Keywords and phrases: Cox regression, finite sample, lasso, oracle inequality, variable selection.

1. Introduction

Since it was introduced by Tibshirani (1996), the lasso regularized method for high-dimensional regression models with sparse coefficients has received a great deal of attention in the literature. Properties of interest for such regression models include the finite sample oracle inequalities. Among the extensive literature of the lasso method, Bunea, Tsybakov, and Wegkamp (2007) and Bickel, Ritov, and Tsybakov (2009) derived the oracle inequalities for prediction risk and estimation error in a general nonparametric regression model including the high-dimensional linear regression as a special example, and van de Geer (2008) provided oracle inequalities for the generalized linear models with Lipschitz loss functions, e.g. logistic regression and classification with hinge loss.

We consider lasso regularized high-dimensional Cox regression. Let T be the survival time and C the censoring time. Suppose we observe a sequence of iid observations (Y_i, Δ_i, X_i) , $i = 1, \dots, n$, where $Y_i = T_i \wedge C_i$, $\Delta_i = I_{\{T_i \leq C_i\}}$, and

*Supported in part by NSF Grant DMS-1007590 and NIH grant R01-AG036802.

X_i are the covariates in \mathcal{X} . Due to largely parallel material, we follow closely the notation in [van de Geer \(2008\)](#). Let

$$\mathcal{F} = \left\{ f_\theta(\cdot) = \sum_{k=1}^m \theta_k \psi_k(\cdot), \theta \in \Theta \right\}.$$

Here Θ is a convex subset of \mathbf{R}^m , and the functions ψ_1, \dots, ψ_m are real-valued basis functions on \mathcal{X} , which are identity functions of corresponding covariates in a standard Cox model.

Consider the following Cox model ([Cox, 1972](#)):

$$\lambda(t|X) = \lambda_0(t)e^{f_\theta(X)},$$

where θ is the parameter of interest and λ_0 is the unknown baseline hazard function. The negative log partial likelihood function for θ becomes

$$l_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \left\{ f_\theta(X_i) - \log \left[\frac{1}{n} \sum_{j=1}^n 1(Y_j \geq Y_i) e^{f_\theta(X_j)} \right] \right\} \Delta_i. \quad (1.1)$$

The corresponding estimator with lasso penalty is denoted by

$$\hat{\theta}_n := \arg \min_{\theta \in \Theta} \{l_n(\theta) + \lambda_n \hat{I}(\theta)\},$$

where $\hat{I}(\theta) := \sum_{k=1}^m \hat{\sigma}_k |\theta_k|$ is the weighted l_1 norm of the vector $\theta \in \mathbf{R}^m$, with random weights $\hat{\sigma}_k := [1/n \sum_{i=1}^n \psi_k^2(X_i)]^{1/2}$.

Clearly the negative log partial likelihood is a sum of non-iid random variables. For ease of theoretical calculation, it is natural to consider the following intermediate function as a “replacement” of the negative log partial likelihood function:

$$\tilde{l}_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \{f_\theta(X_i) - \log \mu(Y_i; f_\theta)\} \Delta_i, \quad (1.2)$$

which has the desirable iid structure, but with an unknown population expectation

$$\mu(t; f_\theta) = E_{X,Y} \left\{ 1(Y \geq t) e^{f_\theta(X)} \right\}.$$

The negative log partial likelihood function (1.1) can then be viewed as a “working” model for the empirical loss function (1.2), and the corresponding loss function becomes

$$\gamma_{f_\theta} = \gamma(f_\theta(X), Y, \Delta) := -\{f_\theta(X) - \log \mu(Y; f_\theta)\} \Delta, \quad (1.3)$$

with expected loss

$$l(\theta) = -E_{Y,\Delta,X} [\{f_\theta(X) - \log \mu(Y; f_\theta)\} \Delta] = P\gamma_{f_\theta}, \quad (1.4)$$

where P denotes the distribution of (Y, Δ, X) . Define the target function \bar{f} by

$$\bar{f} := \arg \min_{f \in \mathbf{F}} P\gamma_f,$$

where $\mathbf{F} \supseteq \mathcal{F}$. For simplicity we will assume that there is a unique minimum as in [van de Geer \(2008\)](#). Uniqueness holds for the regular Cox model when $\mathbf{F} = \mathcal{F}$, see for example, [Andersen and Gill \(1982\)](#). Define the excess risk of f by

$$\mathcal{E}(f) := P\gamma_f - P\gamma_{\bar{f}}.$$

It is desirable to show similar non-asymptotic oracle inequalities for the Cox regression model as in, for example, [van de Geer \(2008\)](#) for generalized linear models. That is, with large probability,

$$\mathcal{E}(f_{\hat{\theta}_n}) \leq \text{const.} \times \min_{\theta \in \Theta} \{\mathcal{E}(f_\theta) + \mathcal{V}_\theta\}.$$

Here \mathcal{V}_θ is called the “estimation error” by [van de Geer \(2008\)](#), which is typically proportional to λ_n^2 times the number of nonzero elements in θ .

Note that the summands in the negative log partial likelihood function (1.1) are not iid, and the intermediate loss function $\gamma(\cdot, Y, \Delta)$ given in (1.3) is not Lipschitz. Hence the conclusion of [van de Geer \(2008\)](#) can not be applied directly. With the Lipschitz condition in [van de Geer \(2008\)](#) replaced by a similar boundedness assumption for regression parameters in [Bühlmann \(2006\)](#), we tackle the problem using pointwise arguments to obtain the oracle bounds of two types of errors: one is between empirical loss (1.2) and expected loss (1.4), and one is between the negative log partial likelihood (1.1) and empirical loss (1.2).

The article is organized as follows. In Section 2, we provide assumptions and additional notation that will be used throughout the paper. In Section 3, following the flow of [van de Geer \(2008\)](#), we first consider the case where the weights $\sigma_k := [E\psi_k^2(X)]^{1/2}$ are fixed, then discuss briefly the case with random weights $\hat{\sigma}_k$.

2. Assumptions

We impose five basic assumptions in this section. Assumptions A, B, and C are identical to the corresponding assumptions in [van de Geer \(2008\)](#). Assumption D has a similar flavor to the assumption (A2) in [Bühlmann \(2006\)](#) for the persistency property of boosting method in high-dimensional linear regression models. Here it replaces the Lipschitz assumption in [van de Geer \(2008\)](#). Assumption E is commonly used for survival models with censored data, see for example, [Andersen and Gill \(1982\)](#).

ASSUMPTION A. $K_m := \max_{1 \leq k \leq m} \{\|\psi_k\|_\infty / \sigma_k\} < \infty$.

ASSUMPTION B. There exists an $\eta > 0$ and strictly convex increasing G , such that for all $\theta \in \Theta$ with $\|f_\theta - \bar{f}\|_\infty \leq \eta$, one has $\mathcal{E}(f_\theta) \geq G(\|f_\theta - \bar{f}\|)$.

ASSUMPTION C. There exists a function $D(\cdot)$ on the subsets of the index set $\{1, \dots, m\}$, such that for all $\mathcal{K} \subset \{1, \dots, m\}$, and for all $\theta \in \Theta$ and $\tilde{\theta} \in \Theta$, we have $\sum_{k \in \mathcal{K}} \sigma_k |\theta_k - \tilde{\theta}_k| \leq \sqrt{D(\mathcal{K})} \|f_\theta - f_{\tilde{\theta}}\|$.

ASSUMPTION D. $L_m := \sup_{\theta \in \Theta} \sum_{k=1}^m |\theta_k| < \infty$.

ASSUMPTION E. The observation time stops at a finite time $\tau > 0$ with $\pi := P(Y \geq \tau) > 0$.

The convex conjugate of function G given in Assumption B is denoted by H such that $uv \leq G(u) + H(v)$. A typical choice of G is quadratic function with some constant C_0 , i.e. $G(u) = u^2/C_0$, see [van de Geer \(2008\)](#).

From Assumptions A, D and E, we have for any $\theta \in \Theta$,

$$e^{|f_\theta(X_i)|} \leq e^{K_m L_m \sigma_{(m)}} := U_m < \infty \quad (2.1)$$

for all i , where $\sigma_{(m)} = \max_{1 \leq k \leq m} \sigma_k$.

Let $I(\theta) := \sum_{k=1}^m \sigma_k |\theta_k|$ be the theoretical l_1 norm of θ , and $\hat{I}(\theta) := \sum_{k=1}^m \hat{\sigma}_k |\theta_k|$ be the empirical l_1 norm. For any θ and $\tilde{\theta}$ in Θ , denote

$$I_1(\theta|\tilde{\theta}) := \sum_{k: \tilde{\theta}_k \neq 0} \sigma_k |\theta_k|, \quad I_2(\theta|\tilde{\theta}) := I(\theta) - I_1(\theta|\tilde{\theta}).$$

Similarly we have corresponding empirical versions,

$$\hat{I}_1(\theta|\tilde{\theta}) := \sum_{k: \tilde{\theta}_k \neq 0} \hat{\sigma}_k |\theta_k|, \quad \hat{I}_2(\theta|\tilde{\theta}) := \hat{I}(\theta) - \hat{I}_1(\theta|\tilde{\theta}).$$

3. Main results

3.1. Non-random normalization weights in the penalty

We show that a similar result to Theorem A.4 of [van de Geer \(2008\)](#) holds for the Cox model. Suppose that $\sigma_k = [E\psi_k^2]^{1/2}$ are known and consider the estimator

$$\hat{\theta}_n := \arg \min_{\theta \in \Theta} \{l_n(\theta) + \lambda_n I(\theta)\}.$$

Denote the empirical probability measure based on the sample $\{(X_i, Y_i, \Delta_i) : i = 1, \dots, n\}$ by P_n . Let $\varepsilon_1, \dots, \varepsilon_n$ be a Rademacher sequence, independent of the training data $(X_1, Y_1, \Delta_1), \dots, (X_n, Y_n, \Delta_n)$. We fix some $\theta^* \in \Theta$ and denote $\mathcal{F}_M := \{f_\theta : \theta \in \Theta, I(\theta - \theta^*) \leq M\}$ for some $M > 0$. For any θ where $I(\theta - \theta^*) \leq M$, denote

$$Z_\theta(M) := |(P_n - P)[\gamma_{f_\theta} - \gamma_{f_{\theta^*}}]| = \left| \left[\tilde{l}_n(\theta) - l(\theta) \right] - \left[\tilde{l}_n(\theta^*) - l(\theta^*) \right] \right|.$$

Note that [van de Geer \(2008\)](#) has considered the supremum of the above $Z_\theta(M)$ over Θ . We find that the pointwise argument is adequate for our purpose because only the lasso estimator is of interest, and that the calculation with $\sup_{f \in \mathcal{F}_M} Z_\theta(M)$ in [van de Geer \(2008\)](#) does not apply to the Cox model due to the lack of Lipschitz property.

Lemma 3.1. *Under Assumptions A, D and E, for all θ satisfying $I(\theta - \theta^*) \leq M$, we have*

$$EZ_\theta(M) \leq \bar{a}_n M,$$

where

$$\bar{a}_n = 4a_n, \quad a_n = \sqrt{\frac{2K_m^2 \log(2m)}{n}} + \frac{K_m \log(2m)}{n}.$$

Proof. By the symmetrization theorem, see e.g. [van der Vaart and Wellner \(1996\)](#) or Theorem A.2 in [van de Geer \(2008\)](#), for a class of only one function we have

$$\begin{aligned} EZ_\theta(M) &\leq 2E \left(\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \{ [f_\theta(X_i) - \log \mu(Y_i; f_\theta)] \Delta_i \right. \right. \\ &\quad \left. \left. - [f_{\theta^*}(X_i) - \log \mu(Y_i; f_{\theta^*})] \Delta_i \} \right| \right) \\ &\leq 2E \left(\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \{ f_\theta(X_i) - f_{\theta^*}(X_i) \} \Delta_i \right| \right) \\ &\quad + 2E \left(\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \{ \log \mu(Y_i; f_\theta) - \log \mu(Y_i; f_{\theta^*}) \} \Delta_i \right| \right) \\ &= A + B. \end{aligned}$$

For A we have

$$A \leq 2 \left(\sum_{k=1}^m \sigma_k |\theta_k - \theta_k^*| \right) E \left(\max_{1 \leq k \leq m} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \Delta_i \psi_k(X_i) / \sigma_k \right| \right).$$

Applying Lemma A.1 in [van de Geer \(2008\)](#) with $\eta_n = K_m$ and $\tau_n^2 = K_m^2$, we obtain

$$E \left(\max_{1 \leq k \leq m} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \Delta_i \frac{\psi_k(X_i)}{\sigma_k} \right| \right) \leq a_n.$$

Thus we have

$$A \leq 2a_n M. \tag{3.1}$$

For B , instead of using the contraction theorem that requires Lipschitz, we use the mean value theorem in the following:

$$\begin{aligned}
& \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \{ \log \mu(Y_i; f_\theta) - \log \mu(Y_i; f_{\theta^*}) \} \Delta_i \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \Delta_i \sum_{k=1}^m \frac{1}{\mu(Y_i; f_{\theta^{**}})} \int_{Y_i}^\infty \int_{\mathcal{X}} (\theta_k - \theta_k^*) \psi_k(x) e^{f_{\theta^{**}}(x)} dP_{X,Y}(x, y) \right| \\
&= \left| \sum_{k=1}^m \sigma_k (\theta_k - \theta_k^*) \frac{1}{n} \sum_{i=1}^n \frac{\varepsilon_i \Delta_i}{\mu(Y_i; f_{\theta^{**}}) \sigma_k} \int_{Y_i}^\infty \int_{\mathcal{X}} \psi_k(x) e^{f_{\theta^{**}}(x)} dP_{X,Y}(x, y) \right| \\
&\leq \left| \sum_{k=1}^m \sigma_k (\theta_k - \theta_k^*) \right| \max_{1 \leq k \leq m} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \Delta_i F_{\theta^{**}}(k, Y_i) \right| \\
&\leq M \max_{1 \leq k \leq m} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \Delta_i F_{\theta^{**}}(k, Y_i) \right|,
\end{aligned}$$

where θ^{**} is between θ and θ^* , and

$$\begin{aligned}
F_{\theta^{**}}(k, t) &= \frac{E[1(Y \geq t) \psi_k(X) e^{f_{\theta^{**}}(X)}]}{\mu(t; f_{\theta^{**}}) \sigma_k} \\
&\leq \frac{(\|\psi_k\|_\infty / \sigma_k) E[1(Y \geq t) e^{f_{\theta^{**}}(X)}]}{\mu(t; f_{\theta^{**}})} \leq K_m. \quad (3.2)
\end{aligned}$$

Since for all i ,

$$\begin{aligned}
& E[\varepsilon_i \Delta_i F_{\theta^{**}}(k, Y_i)] = 0, \quad \|\varepsilon_i \Delta_i F_{\theta^{**}}(k, Y_i)\|_\infty \leq K_m, \text{ and} \\
& \frac{1}{n} \sum_{i=1}^n E[\varepsilon_i \Delta_i F_{\theta^{**}}(k, Y_i)]^2 \leq \frac{1}{n} \sum_{i=1}^n E[F_{\theta^{**}}(k, Y_i)]^2 \leq EK_m^2 = K_m^2,
\end{aligned}$$

following Lemma A.1 in [van de Geer \(2008\)](#), we obtain

$$B \leq 2a_n M. \quad (3.3)$$

Combining (3.1) and (3.3), the upper bound for $EZ_\theta(M)$ is achieved. \square

We now can bound $Z_\theta(M)$ using the Bousquet's concentration theorem provided in [van de Geer \(2008\)](#) as Theorem A.1.

Corollary 3.1. *Under Assumptions A, D and E, for all $M > 0$, $r_1 > 0$ and all θ satisfying $I(\theta - \theta^*) \leq M$, it holds that*

$$P(Z_\theta(M) \geq \bar{\lambda}_{n,0}^A M) \leq \exp(-n \bar{a}_n^2 r_1^2),$$

where

$$\bar{\lambda}_{n,0}^A := \bar{\lambda}_{n,0}^A(r_1) := \bar{a}_n \left(1 + 2r_1 \sqrt{2(K_m^2 + \bar{a}_n K_m)} + \frac{4r_1^2 \bar{a}_n K_m}{3} \right)$$

Proof. Using the triangular inequality and the mean value theorem, we obtain

$$\begin{aligned}
|\gamma_{f_\theta} - \gamma_{f_{\theta^*}}| &\leq |f_\theta(X) - f_{\theta^*}(X)|\Delta + |\log \mu(Y; f_\theta) - \log \mu(Y; f_{\theta^*})|\Delta \\
&\leq \left| \sum_{k=1}^m \sigma_k |\theta_k - \theta_k^*| \frac{\psi_k(X)}{\sigma_k} \right| + |\log \mu(Y; f_\theta) - \log \mu(Y; f_{\theta^*})| \\
&\leq MK_m + \sum_{k=1}^m \sigma_k |\theta_k - \theta_k^*| \cdot \max_{1 \leq k \leq m} |F_{\theta^{**}}(k, Y)| \\
&\leq 2MK_m,
\end{aligned}$$

where θ^{**} is between θ and θ^* , $F_{\theta^{**}}(k, Y)$ is defined in (3.2). So we have

$$\|\gamma_{f_\theta} - \gamma_{f_{\theta^*}}\|_\infty \leq 2MK_m,$$

and

$$P(\gamma_{f_\theta} - \gamma_{f_{\theta^*}})^2 \leq 4M^2 K_m^2.$$

Therefore, in view of Bousquet's concentration theorem and Lemma 3.1, for all $M > 0$ and $r_1 > 0$,

$$\begin{aligned}
P\left(Z_\theta(M) \geq \bar{a}_n M \left(1 + 2r_1 \sqrt{2(K_m^2 + \bar{a}_n K_m)} + \frac{4r_1^2 \bar{a}_n K_m}{3}\right)\right) \\
\leq \exp(-n\bar{a}_n^2 r_1^2).
\end{aligned}$$

□

Now for any θ satisfying $I(\theta - \theta^*) \leq M$, we bound

$$R_\theta(M) := \left| \left[l_n(\theta) - \tilde{l}_n(\theta) \right] - \left[l_n(\theta^*) - \tilde{l}_n(\theta^*) \right] \right|,$$

which is equal to

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n \left| \left[\log \frac{1}{n} \sum_{j=1}^n \frac{1(Y_j \geq Y_i) e^{f_\theta(X_j)}}{\mu(Y_i; f_\theta)} - \log \frac{1}{n} \sum_{j=1}^n \frac{1(Y_j \geq Y_i) e^{f_{\theta^*}(X_j)}}{\mu(Y_i; f_{\theta^*})} \right] \Delta_i \right| \\
&\leq \sup_{0 \leq t \leq \tau} \left| \log \frac{1}{n} \sum_{j=1}^n \frac{1(Y_j \geq t) e^{f_\theta(X_j)}}{\mu(t; f_\theta)} - \log \frac{1}{n} \sum_{j=1}^n \frac{1(Y_j \geq t) e^{f_{\theta^*}(X_j)}}{\mu(t; f_{\theta^*})} \right|.
\end{aligned}$$

By the mean value theorem, we have

$$\begin{aligned}
R_\theta(M) &\leq \sup_{0 \leq t \leq \tau} \left| \sum_{k=1}^m (\theta_k - \theta_k^*) \left\{ \frac{\sum_{j=1}^n 1(Y_j \geq t) e^{f_{\theta^{**}}(X_j)}}{\mu(t; f_{\theta^{**}})} \right\}^{-1} \right. \\
&\quad \left. \left\{ \frac{\sum_{j=1}^n 1(Y_j \geq t) \psi_k(X_j) e^{f_{\theta^{**}}(X_j)}}{\mu(t; f_{\theta^{**}})} - \frac{\sum_{j=1}^n 1(Y_j \geq t) e^{f_{\theta^{**}}(X_j)} E[1(Y \geq t) \psi_k(X) e^{f_{\theta^{**}}(X)}]}{\mu(t; f_{\theta^{**}})^2} \right\} \right| \\
&= \sup_{0 \leq t \leq \tau} \left| \sum_{k=1}^m \sigma_k (\theta_k - \theta_k^*) \left\{ \frac{\sum_{j=1}^n 1(Y_j \geq t) \{\psi_k(X_j)/\sigma_k\} e^{f_{\theta^{**}}(X_j)}}{\sum_{j=1}^n 1(Y_j \geq t) e^{f_{\theta^{**}}(X_j)}} \right. \right. \\
&\quad \left. \left. - \frac{E[1(Y \geq t) \{\psi_k(X)/\sigma_k\} e^{f_{\theta^{**}}(X)}]}{E[1(Y \geq t) e^{f_{\theta^{**}}(X)}]} \right\} \right| \\
&\leq M \sup_{0 \leq t \leq \tau} \left[\frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) e^{f_{\theta^{**}}(X_i)} \right]^{-1} \tag{3.4} \\
&\quad \sup_{0 \leq t \leq \tau} \left\{ \max_{1 \leq k \leq m} \left| \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) \{\psi_k(X_i)/\sigma_k\} e^{f_{\theta^{**}}(X_i)} \right. \right. \\
&\quad \left. \left. - E[1(Y \geq t) \{\psi_k(X)/\sigma_k\} e^{f_{\theta^{**}}(X)}] \right| \right. \\
&\quad \left. + K_m \left| \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) e^{f_{\theta^{**}}(X_i)} - E[1(Y \geq t) e^{f_{\theta^{**}}(X)}] \right| \right\},
\end{aligned}$$

where θ^{**} is between θ and θ^* , and by (2.1) we have

$$\sup_{0 \leq t \leq \tau} \left[\frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) e^{f_{\theta^{**}}(X_i)} \right]^{-1} \leq U_m \left[\frac{1}{n} \sum_{i=1}^n 1(Y_i \geq \tau) \right]^{-1}. \tag{3.5}$$

Lemma 3.2. *Under Assumption E, we have*

$$P \left(\frac{1}{n} \sum_{i=1}^n 1(Y_i \geq \tau) \leq \frac{\pi}{2} \right) \leq 2e^{-n\pi^2/2}.$$

Proof. This is obtained directly from Massart (1990) by taking $r = \pi\sqrt{n}/2$ in the following:

$$\begin{aligned}
P \left(\frac{1}{n} \sum_{i=1}^n 1(Y_i \geq \tau) \leq \frac{\pi}{2} \right) &\leq P \left(\sup_{0 \leq t \leq \tau} \sqrt{n} \left| \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) - \pi \right| \geq r \right) \\
&\leq 2e^{-2r^2}.
\end{aligned}$$

□

Lemma 3.3. *Under Assumptions A, D and E, for all θ we have*

$$\begin{aligned} P\left(\sup_{0 \leq t \leq \tau} \left| \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) e^{f_\theta(X_i)} - \mu(t; f_\theta) \right| \geq U_m \bar{a}_n r_1\right) \\ \leq \frac{1}{5} W^2 e^{-n \bar{a}_n^2 r_1^2}, \end{aligned} \quad (3.6)$$

where W is a constant that only depends on $K = \sqrt{2}$.

Proof. For a class of functions indexed by t , $\mathcal{F} = \{1(y \geq t) e^{f_\theta(x)} / U_m : t \in [0, \tau], y \in \mathbf{R}, e^{f_\theta(x)} \leq U_m\}$, we calculate its bracketing number. For any $\epsilon > 0$, let t_i be the i -th $\lceil 1/\epsilon \rceil$ quantile of Y , i.e.,

$$P(Y \leq t_i) = i\epsilon, \quad i = 1, \dots, \lceil 1/\epsilon \rceil - 1,$$

where $\lceil x \rceil$ is the smallest integer that is greater than or equal to x . Furthermore, denote $t_0 = 0$ and $t_{\lceil 1/\epsilon \rceil} = +\infty$. For $i = 1, \dots, \lceil 1/\epsilon \rceil$, define brackets $[L_i, U_i]$ with

$$L_i(x, y) = 1(y \geq t_i) e^{f_\theta(x)} / U_m, \quad U_i(x, y) = 1(y > t_{i-1}) e^{f_\theta(x)} / U_m$$

such that $L_i(x, y) \leq 1(y \geq t) e^{f_\theta(x)} / U_m \leq U_i(x, y)$ when $t_{i-1} < t \leq t_i$. Since

$$\begin{aligned} \{E[U_i - L_i]^2\}^{1/2} &\leq \left\{ E \left[\frac{e^{f_\theta(X)}}{U_m} \{1(Y \geq t_i) - 1(Y > t_{i-1})\} \right]^2 \right\}^{1/2} \\ &\leq \{P(t_{i-1} < Y \leq t_i)\}^{1/2} = \sqrt{\epsilon}, \end{aligned}$$

we have $N_{[]}(\sqrt{\epsilon}, \mathcal{F}, L_2) \leq 2/\epsilon$, which yields

$$N_{[]}(\epsilon, \mathcal{F}, L_2) \leq \frac{2}{\epsilon^2} = \left(\frac{K}{\epsilon} \right)^2,$$

where $K = \sqrt{2}$. Thus, from Theorem 2.14.9 in [van der Vaart and Wellner \(1996\)](#), we have for any $r > 0$,

$$\begin{aligned} P\left(\sqrt{n} \sup_{0 \leq t \leq \tau} \left| \frac{1}{n} \sum_{i=1}^n \frac{1(Y_i \geq t) e^{f_\theta(X_i)}}{U_m} - \frac{\mu(t; f_\theta)}{U_m} \right| \geq r\right) &\leq \frac{1}{2} W^2 r^2 e^{-2r^2} \\ &\leq \frac{1}{5} W^2 e^{-r^2}, \end{aligned}$$

where W is a constant that only depends on K . Note that $r^2 e^{-r^2}$ is bounded by e^{-1} . Let $r = \sqrt{n} \bar{a}_n r_1$, we obtain (3.6). \square

Lemma 3.4. *Under Assumptions A, D and E, for all θ we have*

$$\begin{aligned} & P\left(\sup_{0 \leq t \leq \tau} \max_{0 \leq k \leq m} \left| \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) \frac{\psi_k(X_i)}{\sigma_k} e^{f_\theta(X_i)} \right. \right. \\ & \quad \left. \left. - E \left[1(Y \geq t) \frac{\psi_k(X)}{\sigma_k} e^{f_\theta(X)} \right] \right| \geq K_m U_m \left[\bar{a}_n r_1 + \sqrt{\frac{\log(2m)}{n}} \right] \right) \\ & \leq \frac{1}{10} W^2 e^{-n \bar{a}_n^2 r_1^2}. \end{aligned} \quad (3.7)$$

Proof. Consider the classes of functions indexed by t ,

$$\begin{aligned} \mathcal{G}^k &= \{ 1(y \geq t) e^{f_\theta(x)} \psi_k(x) / (\sigma_k K_m U_m) : t \in [0, \tau], y \in \mathbf{R}, \\ & \quad |e^{f_\theta(x)} \psi_k(x) / \sigma_k| \leq K_m U_m \}, \quad k = 1, \dots, m. \end{aligned}$$

Using the same argument in the proof of Lemma 3.3, we have

$$N_{[]}(\varepsilon, \mathcal{G}^k, L_2) \leq \left(\frac{K}{\varepsilon} \right)^2,$$

where $K = \sqrt{2}$, and then for any $r > 0$,

$$\begin{aligned} & P\left(\sqrt{n} \sup_{0 \leq t \leq \tau} \left| \frac{1}{n} \sum_{i=1}^n \frac{1(Y_i \geq t) e^{f_\theta(X_i)} \psi_k(X_i)}{\sigma_k K_m U_m} \right. \right. \\ & \quad \left. \left. - E \left[\frac{1(Y \geq t) e^{f_\theta(X)} \psi_k(X)}{\sigma_k K_m U_m} \right] \right| \geq r \right) \leq \frac{1}{5} W^2 e^{-r^2}. \end{aligned}$$

Thus we have

$$\begin{aligned} & P\left(\sqrt{n} \sup_{0 \leq t \leq \tau} \max_{0 \leq k \leq m} \left| \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) e^{f_\theta(X_i)} \psi_k(X_i) / (\sigma_k U_m K_m) \right. \right. \\ & \quad \left. \left. - E \left[1(Y \geq t) e^{f_\theta(X)} \psi_k(X) / (\sigma_k U_m K_m) \right] \right| \geq r \right) \\ & \leq P\left(\bigcup_{k=1}^m \sqrt{n} \sup_{0 \leq t \leq \tau} \left| \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) e^{f_\theta(X_i)} \psi_k(X_i) / (\sigma_k U_m K_m) \right. \right. \\ & \quad \left. \left. - E \left[1(Y \geq t) e^{f_\theta(X)} \psi_k(X) / (\sigma_k U_m K_m) \right] \right| \geq r \right) \\ & \leq m P\left(\sqrt{n} \sup_{0 \leq t \leq \tau} \left| \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) e^{f_\theta(X_i)} \psi_k(X_i) / (\sigma_k U_m K_m) \right. \right. \\ & \quad \left. \left. - E \left[1(Y \geq t) e^{f_\theta(X)} \psi_k(X) / (\sigma_k U_m K_m) \right] \right| \geq r \right) \\ & \leq \frac{m}{5} W^2 e^{-r^2} = \frac{1}{10} W^2 e^{\log(2m) - r^2}. \end{aligned}$$

Let $\log(2m) - r^2 = -n\bar{a}_n^2 r_1^2$, i.e. $r = \sqrt{n\bar{a}_n^2 r_1^2 + \log(2m)}$. Since

$$\sqrt{\bar{a}_n^2 r_1^2 + \frac{\log(2m)}{n}} \leq \bar{a}_n r_1 + \sqrt{\frac{\log(2m)}{n}},$$

we obtain (3.7). \square

Corollary 3.2. *Under Assumptions A, D and E, for all $M > 0$ and all θ that satisfies $I(\theta - \theta^*) \leq M$, we have*

$$P(R_\theta(M) \geq \bar{\lambda}_{n,0}^B M) \leq 2 \exp(-n\pi^2/2) + \frac{3}{10} W^2 \exp(-n\bar{a}_n^2 r_1^2), \quad (3.8)$$

where

$$\bar{\lambda}_{n,0}^B = \frac{2K_m U_m^2}{\pi} \left(2\bar{a}_n r_1 + \sqrt{\frac{\log(2m)}{n}} \right).$$

Proof. From inequalities (3.4) and (3.5) we have

$$P(R_\theta(M) \leq \bar{\lambda}_{n,0}^B \cdot M) \geq P(E_1^c \cap E_2^c \cap E_3^c),$$

where the events E_1 , E_2 and E_3 are defined in the following:

$$\begin{aligned} E_1 &= \left\{ \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq \tau) \leq \pi/2 \right\}, \\ E_2 &= \left\{ \sup_{0 \leq t \leq \tau} \left| \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) e^{f_{\theta^{**}}(X_i)} - \mu(t; f_{\theta^{**}}) \right| \geq U_m \bar{a}_n r_1 \right\}, \\ E_3 &= \left\{ \max_{0 \leq k \leq m} \sup_{0 \leq t \leq \tau} \left| \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) \frac{\psi_k(X_i)}{\sigma_k} e^{f_{\theta^{**}}(X_i)} \right. \right. \\ &\quad \left. \left. - E \left[1(Y \geq t) \frac{\psi_k(X)}{\sigma_k} e^{f_{\theta^{**}}(X)} \right] \right| \geq K_m U_m (\bar{a}_n r_1 + \sqrt{\frac{\log(2m)}{n}}) \right\}. \end{aligned}$$

Thus

$$P(R_\theta(M) \geq \bar{\lambda}_{n,0}^B \cdot M) \leq P(E_1^c) + P(E_2^c) + P(E_3^c),$$

and the result follows from Lemmas 3.2, 3.3 and 3.4. \square

We now show oracle bounds for the lasso estimator $\hat{\theta}_n$ under Assumptions A-E following van de Geer (2008), but using pointwise arguments. Let

$$\bar{\lambda}_{n,0} = \bar{\lambda}_{n,0}^A + \bar{\lambda}_{n,0}^B. \quad (3.9)$$

Take $b > 0$, $d > 1$, and

$$d_b := d \left(\frac{b+d}{(d-1)b} \vee 1 \right).$$

Let $D_\theta := D(\{k : \theta_k \neq 0, k = 1, \dots, m\})$ be the number of nonzero θ_k 's, where $D(\cdot)$ is given in Assumption C. Define

$$\begin{aligned}
(A1) \quad \lambda_n &:= (1+b)\bar{\lambda}_{n,0}, \\
(A2) \quad \mathcal{V}_\theta &:= 2\delta H\left(\frac{2\lambda_n\sqrt{D_\theta}}{\delta}\right), \text{ where } 0 < \delta < 1, \\
(A3) \quad \theta_n^* &:= \operatorname{argmin}_{\theta \in \Theta} \{\mathcal{E}(f_\theta) + \mathcal{V}_\theta\}, \\
(A4) \quad \epsilon_n^* &:= (1+\delta)\mathcal{E}(f_{\theta_n^*}) + \mathcal{V}_{\theta_n^*}, \\
(A5) \quad \zeta_n^* &:= \frac{\epsilon_n^*}{\lambda_{n,0}}, \\
(A6) \quad \theta(\epsilon_n^*) &:= \operatorname{argmin}_{\theta \in \Theta, I(\theta - \theta_n^*) \leq d_b \zeta_n^*/b} \{\delta \mathcal{E}(f_\theta) - 2\lambda_n I_1(\theta - \theta_n^* | \theta_n^*)\}.
\end{aligned}$$

We also impose the following conditions:

CONDITION I(b, δ). $\|f_{\theta_n^*} - \bar{f}\|_\infty \leq \eta$.

CONDITION II(b, δ, d). $\|f_{\theta(\epsilon_n^*)} - \bar{f}\|_\infty \leq \eta$.

In both conditions, η is given in Assumption B.

Lemma 3.5. *Suppose Conditions I(b, δ) and II(b, δ, d) are met. For all $\theta \in \Theta$ with $I(\theta - \theta_n^*) \leq d_b \zeta_n^*/b$, it holds that*

$$2\lambda_n I_1(\theta - \theta_n^*) \leq \delta \mathcal{E}(f_\theta) + \epsilon_n^* - \mathcal{E}(f_{\theta_n^*}).$$

Proof. The proof is exactly the same as that of Lemma A.4 in [van de Geer \(2008\)](#), with λ_n defined in (3.9). \square

Lemma 3.6. *Suppose Conditions I(b, δ) and II(b, δ, d) are met. Consider any random $\tilde{\theta} \in \Theta$ with $l_n(\tilde{\theta}) + \lambda_n I(\tilde{\theta}) \leq l_n(\theta_n^*) + \lambda_n I(\theta_n^*)$. Let $1 < d_0 \leq d_b$. It holds that*

$$\begin{aligned}
P\left(I(\tilde{\theta} - \theta_n^*) \leq d_0 \frac{\zeta_n^*}{b}\right) &\leq P\left(I(\tilde{\theta} - \theta_n^*) \leq \left(\frac{d_0 + b}{1 + b}\right) \frac{\zeta_n^*}{b}\right) \\
&\quad + \left(1 + \frac{3}{10}W^2\right) \exp(-n\bar{a}_n^2 r_1^2) + 2 \exp(-n\pi^2/2).
\end{aligned}$$

Proof. The idea is similar to the proof of Lemma A.5 in [van de Geer \(2008\)](#). Let $\tilde{\mathcal{E}} = \mathcal{E}(f_{\tilde{\theta}})$ and $\mathcal{E}^* = \mathcal{E}(f_{\theta_n^*})$. We will use short notation: $I_1(\theta) = I_1(\theta | \theta_n^*)$ and $I_2(\theta) = I_2(\theta | \theta_n^*)$. Since $l_n(\tilde{\theta}) + \lambda_n I(\tilde{\theta}) \leq l_n(\theta_n^*) + \lambda_n I(\theta_n^*)$, on the set where $I(\tilde{\theta} - \theta_n^*) \leq d_0 \zeta_n^*/b$ and $Z_{\tilde{\theta}}(d_0 \zeta_n^*/b) \leq d_0 \zeta_n^*/b \cdot \bar{\lambda}_{n,0}^A$, we have

$$\begin{aligned}
R_{\tilde{\theta}}(d_0 \zeta_n^*/b) &\geq [l_n(\theta_n^*) + \lambda_n I(\theta_n^*)] - [l_n(\tilde{\theta}) + \lambda_n I(\tilde{\theta})] - \lambda_n I(\theta_n^*) + \lambda_n I(\tilde{\theta}) \\
&\quad - [\tilde{l}_n(\theta_n^*) - \tilde{l}_n(\tilde{\theta})] \\
&\geq -\lambda_n I(\theta_n^*) + \lambda_n I(\tilde{\theta}) - [\tilde{l}_n(\theta_n^*) - \tilde{l}_n(\tilde{\theta})] \\
&\geq -\lambda_n I(\theta_n^*) + \lambda_n I(\tilde{\theta}) - [l(\theta_n^*) - l(\tilde{\theta})] - d_0 \zeta_n^*/b \cdot \bar{\lambda}_{n,0}^A \\
&\geq -\lambda_n I(\theta_n^*) + \lambda_n I(\tilde{\theta}) - \mathcal{E}^* + \tilde{\mathcal{E}} - d_0 \bar{\lambda}_{n,0}^A \zeta_n^*/b.
\end{aligned} \tag{3.10}$$

By (3.8) we know that $R_{\tilde{\theta}}(d_0\zeta_n^*/b)$ is bounded by $d_0\bar{\lambda}_{n,0}^B\zeta_n^*/b$ with probability at least $1 - \frac{3}{10}W^2 \exp(-n\bar{a}_n^2 r_1^2) - 2 \exp(-n\pi^2/2)$, then we have

$$\tilde{\mathcal{E}} + \lambda_n I(\tilde{\theta}) \leq \bar{\lambda}_{n,0}^B d_0 \zeta_n^*/b + \mathcal{E}^* + \lambda_n I(\theta_n^*) + \bar{\lambda}_{n,0}^A d_0 \zeta_n^*/b.$$

Since $I(\tilde{\theta}) = I_1(\tilde{\theta}) + I_2(\tilde{\theta})$ and $I(\theta_n^*) = I_1(\theta_n^*)$, using the triangular inequality, we obtain

$$\begin{aligned} \tilde{\mathcal{E}} + (1+b)\bar{\lambda}_{n,0}I_2(\tilde{\theta}) &\leq \bar{\lambda}_{n,0}d_0\zeta_n^*/b + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0}I_1(\theta_n^*) - (1+b)\bar{\lambda}_{n,0}I_1(\tilde{\theta}) \\ &\leq \bar{\lambda}_{n,0}d_0\zeta_n^*/b + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0}I_1(\tilde{\theta} - \theta_n^*). \end{aligned} \quad (3.11)$$

The remaining of the proof follows exactly the same as the corresponding part of the proof of Lemma A.5 in van de Geer (2008). \square

Corollary 3.3. *Suppose Conditions I(b, δ) and II(b, δ , d) are met. Consider any random $\tilde{\theta} \in \Theta$ with $l_n(\tilde{\theta}) + \lambda_n I(\tilde{\theta}) \leq l_n(\theta_n^*) + \lambda_n I(\theta_n^*)$. Let $1 < d_0 \leq d_b$. It holds that*

$$\begin{aligned} P\left(I(\tilde{\theta} - \theta_n^*) \leq d_0 \frac{\zeta_n^*}{b}\right) &\leq P\left(I(\tilde{\theta} - \theta_n^*) \leq [1 + (d_0 - 1)(1+b)^{-N}] \frac{\zeta_n^*}{b}\right) \\ &\quad + N \left\{ \left(1 + \frac{3}{10}W^2\right) \exp(-n\bar{a}_n^2 r_1^2) + 2 \exp(-n\pi^2/2) \right\}. \end{aligned}$$

Proof. Repeat Lemma 3.6 N times. \square

Lemma 3.7. *Suppose Conditions I(b, δ) and II(b, δ , d) are met. Define*

$$\tilde{\theta}_s = s\hat{\theta}_n + (1-s)\theta_n^*,$$

where

$$s = \frac{d\zeta_n^*}{d\zeta_n^* + bI(\hat{\theta}_n - \theta_n^*)}.$$

Then for any integer N , with probability at least

$$1 - N \left\{ \left(1 + \frac{3}{10}W^2\right) \exp(-n\bar{a}_n^2 r_1^2) + 2 \exp(-n\pi^2/2) \right\},$$

we have

$$I(\tilde{\theta}_s - \theta_n^*) \leq (1 + (d-1)(1+b)^{-N}) \frac{\zeta_n^*}{b}.$$

Proof. Since the negative log partial likelihood $l_n(\theta)$ and the lasso penalty are both convex with respect to θ , applying Corollary 3.3, we obtain the above inequality. \square

Lemma 3.8. *Suppose Conditions $I(b, \delta)$ and $II(b, \delta, d)$ are met. Let $N_1 \in \mathbf{N} := \{1, 2, \dots\}$ and $N_2 \in \mathbf{N} \cup \{0\}$. Define $\delta_1 = (1+b)^{-N_1}$ and $\delta_2 = (1+b)^{-N_2}$. For any n , with probability at least*

$$1 - (N_1 + N_2) \left\{ \left(1 + \frac{3}{10} W^2 \right) \exp(-n \bar{a}_n^2 r_1^2) + 2 \exp(-n \pi^2 / 2) \right\},$$

we have

$$I(\hat{\theta}_n - \theta_n^*) \leq d(\delta_1, \delta_2) \frac{\zeta_n^*}{b},$$

where

$$d(\delta_1, \delta_2) = 1 + \frac{1 + (d^2 - 1)\delta_1}{(d - 1)(1 - \delta_1)} \delta_2.$$

Proof. The proof is exactly the same as that of Lemma A.7 in [van de Geer \(2008\)](#), with a slightly different probability bound. \square

We now provide the major theorem of the oracle inequalities for the Cox model lasso estimator.

Theorem 3.1. *Suppose Assumptions A-E and Conditions $I(b, \delta)$ and $II(b, \delta, d)$ are met. Let*

$$\Delta(b, \delta, \delta_1, \delta_2) := d(\delta_1, \delta_2) \frac{1 - \delta^2}{\delta b} \vee 1.$$

We have with probability at least

$$1 - \left\{ \log_{1+b} \frac{(1+b)^2 \Delta(b, \delta, \delta_1, \delta_2)}{\delta_1 \delta_2} \right\} \left\{ \left(1 + \frac{3}{10} W^2 \right) \exp(-n \bar{a}_n^2 r_1^2) + 2 \exp(-n \pi^2 / 2) \right\}$$

that

$$\mathcal{E}(f_{\hat{\theta}_n}) \leq \frac{1}{1 - \delta} \epsilon_n^*,$$

and moreover,

$$I(\hat{\theta}_n - \theta_n^*) \leq d(\delta_1, \delta_2) \frac{\zeta_n^*}{b}.$$

Proof. The proof follows the same ideas in the proof of Theorem A.4 in [van de Geer \(2008\)](#), with exceptions of pointwise arguments and slightly different probability bounds. Since this is the major result, to be self-contained, we provide a detailed proof here despite the amount of overlaps.

Similar to [van de Geer \(2008\)](#), we define $\hat{\mathcal{E}} := \mathcal{E}(f_{\hat{\theta}_n})$ and $\mathcal{E}^* := \mathcal{E}(f_{\theta_n^*})$; use the notation $I_1(\theta) := I_1(\theta | \theta_n^*)$ and $I_2(\theta) := I_2(\theta | \theta_n^*)$; set

$$c := \frac{\delta b}{1 - \delta^2};$$

and consider the cases (a) $c < d(\delta_1, \delta_2)$ and (b) $c \geq d(\delta_1, \delta_2)$.

(a) Consider $c < d(\delta_1, \delta_2)$. Let J be an integer satisfying $(1+b)^{J-1}c \leq d(\delta_1, \delta_2)$ and $(1+b)^J c > d(\delta_1, \delta_2)$. We consider the cases (a1) $c\zeta_n^*/b < I(\hat{\theta}_n - \theta_n^*) \leq d(\delta_1, \delta_2)\zeta_n^*/b$ and (a2) $I(\hat{\theta}_n - \theta_n^*) \leq c\zeta_n^*/b$.
 (a1) If $c\zeta_n^*/b < I(\hat{\theta}_n - \theta_n^*) \leq d(\delta_1, \delta_2)\zeta_n^*/b$, then

$$(1+b)^{j-1}c\frac{\zeta_n^*}{b} < I(\hat{\theta}_n - \theta_n^*) \leq (1+b)^j c\frac{\zeta_n^*}{b}$$

for some $j \in \{1, \dots, J\}$. Let

$$d_0 = c(1+b)^{j-1} \leq d(\delta_1, \delta_2) \leq d_b.$$

From Corollary 3.1, with probability at least $1 - \exp(-n\bar{a}_n^2 r_1^2)$ we have $Z_{\hat{\theta}_n}((1+b)d_0\zeta_n^*/b) \leq (1+b)d_0\bar{\lambda}_{n,0}^A\zeta_n^*/b$. Since $l_n(\hat{\theta}_n) + \lambda_n I(\hat{\theta}_n) \leq l_n(\theta_n^*) + \lambda_n I(\theta_n^*)$, from equation (3.10), we have

$$\hat{\mathcal{E}} + \lambda_n I(\hat{\theta}_n) \leq R_{\hat{\theta}_n} \left((1+b)d_0\frac{\zeta_n^*}{b} \right) + \mathcal{E}^* + \lambda_n I(\theta_n^*) + (1+b)\bar{\lambda}_{n,0}^A d_0\frac{\zeta_n^*}{b}.$$

By (3.8), $R_{\hat{\theta}_n}((1+b)d_0\zeta_n^*/b)$ is bounded by $(1+b)\bar{\lambda}_{n,0}^B d_0\zeta_n^*/b$ with probability at least

$$1 - \frac{3}{10}W^2 \exp(-n\bar{a}_n^2 r_1^2) - 2 \exp(-n\pi^2/2),$$

then we have

$$\begin{aligned} \hat{\mathcal{E}} + (1+b)\bar{\lambda}_{n,0} I(\hat{\theta}_n) &\leq (1+b)\bar{\lambda}_{n,0}^B d_0\frac{\zeta_n^*}{b} + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0} I(\theta_n^*) \\ &\quad + (1+b)\bar{\lambda}_{n,0}^A d_0\frac{\zeta_n^*}{b} \\ &\leq (1+b)\bar{\lambda}_{n,0} I(\hat{\theta}_n - \theta_n^*) + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0} I(\theta_n^*). \end{aligned}$$

Since $I(\hat{\theta}_n) = I_1(\hat{\theta}_n) + I_2(\hat{\theta}_n)$, $I(\hat{\theta}_n - \theta_n^*) = I_1(\hat{\theta}_n - \theta_n^*) + I_2(\hat{\theta}_n)$, and $I(\theta_n^*) = I_1(\theta_n^*)$, by triangular inequality we obtain

$$\hat{\mathcal{E}} \leq 2(1+b)\bar{\lambda}_{n,0} I_1(\hat{\theta}_n - \theta_n^*) + \mathcal{E}^*.$$

From Lemma 3.5,

$$\hat{\mathcal{E}} \leq \delta \hat{\mathcal{E}} + \epsilon_n^* - \mathcal{E}^* + \mathcal{E}^* = \delta \hat{\mathcal{E}} + \epsilon_n^*.$$

Hence,

$$\hat{\mathcal{E}} \leq \frac{1}{1-\delta} \epsilon_n^*.$$

(a2) If $I(\hat{\theta}_n - \theta_n^*) \leq c\zeta_n^*/b$, from equation (3.11) with $d_0 = c$, with probability at least

$$1 - \left\{ \left(1 + \frac{3}{10}W^2 \right) \exp(-n\bar{a}_n^2 r_1^2) + 2 \exp(-n\pi^2/2) \right\},$$

we have

$$\hat{\mathcal{E}} + (1+b)\bar{\lambda}_{n,0}I(\hat{\theta}_n) \leq \frac{\delta}{1-\delta^2}\bar{\lambda}_{n,0}\zeta_n^* + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0}I(\theta_n^*).$$

By triangular inequality, Lemma 3.5 and (A4),

$$\begin{aligned} \hat{\mathcal{E}} &\leq \frac{\delta}{1-\delta^2}\bar{\lambda}_{n,0}\zeta_n^* + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0}I_1(\hat{\theta}_n - \theta_n^*) \\ &\leq \frac{\delta}{1-\delta^2}\bar{\lambda}_{n,0}\frac{\epsilon_n^*}{\bar{\lambda}_{n,0}} + \mathcal{E}^* + \frac{\delta}{2}\hat{\mathcal{E}} + \frac{1}{2}\epsilon_n^* - \frac{1}{2}\mathcal{E}^* \\ &= \left(\frac{\delta}{1-\delta^2} + \frac{1}{2}\right)\epsilon_n^* + \frac{1}{2}\mathcal{E}^* + \frac{\delta}{2}\hat{\mathcal{E}} \\ &\leq \left(\frac{\delta}{1-\delta^2} + \frac{1}{2}\right)\epsilon_n^* + \frac{1}{2(1+\delta)}\epsilon_n^* + \frac{\delta}{2}\hat{\mathcal{E}}. \end{aligned}$$

Hence,

$$\hat{\mathcal{E}} \leq \frac{2}{2-\delta} \left[\frac{\delta}{1-\delta^2} + \frac{1}{2} + \frac{1}{2(1+\delta)} \right] \epsilon_n^* = \frac{1}{1-\delta}\epsilon_n^*.$$

Furthermore, by Lemma 3.8, we have with probability at least

$$1 - (N_1 + N_2) \left\{ \left(1 + \frac{3}{10}W^2\right) \exp(-n\bar{a}_n^2 r_1^2) + 2 \exp(-n\pi^2/2) \right\}$$

that

$$I(\hat{\theta}_n - \theta_n^*) \leq d(\delta_1, \delta_2) \frac{\zeta_n^*}{b},$$

where

$$N_1 = \log_{1+b} \left(\frac{1}{\delta_1} \right), \quad N_2 = \log_{1+b} \left(\frac{1}{\delta_2} \right).$$

(b) Consider $c \geq d(\delta_1, \delta_2)$. On the set where $I(\hat{\theta}_n - \theta_n^*) \leq d(\delta_1, \delta_2)\zeta_n^*/b$, from equation (3.11) we have with probability at least

$$1 - \left\{ \left(1 + \frac{3}{10}W^2\right) \exp(-n\bar{a}_n^2 r_1^2) + 2 \exp(-n\pi^2/2) \right\}$$

that

$$\begin{aligned} \hat{\mathcal{E}} + (1+b)\bar{\lambda}_{n,0}I(\hat{\theta}_n) &\leq \bar{\lambda}_{n,0}d(\delta_1, \delta_2) \frac{\zeta_n^*}{b} + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0}I(\theta_n^*) \\ &\leq \frac{\delta}{1-\delta^2}\bar{\lambda}_{n,0}\zeta_n^* + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0}I(\theta_n^*), \end{aligned}$$

which is the same as (a2) and leads to the same result.

To summarize, let

$$A = \left\{ \hat{\mathcal{E}} \leq \frac{1}{1-\delta}\epsilon_n^* \right\}, \quad B = \left\{ I(\hat{\theta}_n - \theta_n^*) \leq d(\delta_1, \delta_2) \frac{\zeta_n^*}{b} \right\}.$$

Note that

$$J + 1 \leq \log_{1+b} \left(\frac{(1+b)^2 d(\delta_1, \delta_2)}{c} \right).$$

Under case (a), we have

$$\begin{aligned} P(A \cap B) &= P(a1) - P(A^c \cap a1) + P(a2) - P(A^c \cap a2) \\ &\geq P(a1) - J \left\{ \left(1 + \frac{3}{10} W^2 \right) \exp(-n\bar{a}_n^2 r_1^2) + 2 \exp(-n\pi^2/2) \right\} \\ &\quad + P(a2) - \left\{ \left(1 + \frac{3}{10} W^2 \right) \exp(-n\bar{a}_n^2 r_1^2) + 2 \exp(-n\pi^2/2) \right\} \\ &= P(B) - (J + 1) \left\{ \left(1 + \frac{3}{10} W^2 \right) \exp(-n\bar{a}_n^2 r_1^2) + 2 \exp(-n\pi^2/2) \right\} \\ &\geq 1 - (N_1 + N_2 + J + 1) \left\{ \left(1 + \frac{3}{10} W^2 \right) \exp(-n\bar{a}_n^2 r_1^2) \right. \\ &\quad \left. + 2 \exp(-n\pi^2/2) \right\} \\ &\geq 1 - \log_{1+b} \left\{ \frac{(1+b)^2}{\delta_1 \delta_2} \cdot \frac{d(\delta_1, \delta_2)(1 - \delta^2)}{\delta b} \right\} \\ &\quad \left\{ \left(1 + \frac{3}{10} W^2 \right) \exp(-n\bar{a}_n^2 r_1^2) + 2 \exp(-n\pi^2/2) \right\}. \end{aligned}$$

Under case (b),

$$\begin{aligned} P(A \cap B) &= P(B) - P(A^c \cap B) \\ &\geq P(B) - \left\{ \left(1 + \frac{3}{10} W^2 \right) \exp(-n\bar{a}_n^2 r_1^2) + 2 \exp(-n\pi^2/2) \right\} \\ &\geq 1 - (N_1 + N_2 + 2) \left\{ \left(1 + \frac{3}{10} W^2 \right) \exp(-n\bar{a}_n^2 r_1^2) \right. \\ &\quad \left. + 2 \exp(-n\pi^2/2) \right\} \\ &= 1 - \log_{1+b} \left\{ \frac{(1+b)^2}{\delta_1 \delta_2} \right\} \\ &\quad \left\{ \left(1 + \frac{3}{10} W^2 \right) \exp(-n\bar{a}_n^2 r_1^2) + 2 \exp(-n\pi^2/2) \right\}. \end{aligned}$$

We thus obtain the desired result. \square

3.2. Random normalization weights in the penalty

The case with random weights can be argued in the exactly the same way as that in [van de Geer \(2008\)](#), for which the same tail probability given in Lemma A.9 of [van de Geer \(2008\)](#) is added to the probability bound in Theorem 3.1 under the same set of conditions for Theorem A.5 in [van de Geer \(2008\)](#). Thus details are omitted.

References

- ANDERSEN, P. K. AND GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10** 1100–1120.
- BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. AND WELLNER, J. A. (1993). Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins Univ. Press
- BICKEL, P., RITOV, Y. AND TSYBAKOV, A. (2009) Simultaneous Analysis of Lasso and Dantzig Selector. *Ann. Statist.*, **37**, 1705–1732.
- BÜHLMANN, P. (2006) Boosting For High-dimensional Linear Models. *Ann. Statist.*, **34**, 559–583.
- BUNEA, F., TSYBAKOV, A. B. AND WEGKAMP, M. H. (2007) Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, **1**, 169–194.
- COX, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220.
- MASSART, P. (1990) The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, **18**, 1269–1283.
- TARIGAN, B. AND VAN DE GEER, S. (2006) Classifiers of support vector machine type with l_1 complexity regularization. *Bernoulli*, **12**, 1045–1076.
- TIBSHIRANI, R. (1996) Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc., Ser. B*, **58**, 267–288.
- VAN DER VAART AND WELLNER, J. (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics*. Wiley, New York.
- VAN DE GEER, S. (2008) High-dimensional generalized linear models and the lasso. *Ann. Statist.*, **36**, 614–645.